Policy Brief No. 206 — August 2025

A Call to Address Anthropomorphic AI Threats to Freedom of Thought

Abel Wajnerman Paz

Key Points

→ Bonding chatbots simulate emotionally meaningful relationships with users, encouraging anthropomorphization and emotional attachment.

→ Socially vulnerable users, who rely on bonding chatbots as their primary source of social connection, are particularly exposed to emotional manipulation, which may undermine their ability to freely flourish through human relationships, an essential aspect of freedom of thought.

→ Current classification of bonding chatbots as "well-being" tools rather than medical products limits regulatory oversight.

→ A medical model for bonding chatbots would provide stronger safeguards, requiring professional supervision and therapeutic framing.

Introduction: A New Threat to Freedom of Thought

Social isolation has been a critical issue for many countries during the past several years (Shaer 2024) and, inspired by a premature techno-optimism, artificial intelligence (AI) is being discussed and implemented as a potential solution (Samuel 2025). A type of conversational AI, referred to in this policy brief as "bonding chatbots," is designed to embody a personal relationship of the user, such as a friend, a sibling or a romantic partner. These artificial bonds are available 24/7 to attend to our emotional needs and reduce our perceived loneliness. However, the personal and societal damage that this technology can cause is already tangible. For instance, last year a case filed against Character.AI (one of the companies developing bonding chatbots) included allegations of a child being encouraged by a chatbot to self-harm or to murder their parents (Tiku 2024). This is not an isolated incident. Last year, the mother of Sewell

About the Author

Abel Wajnerman Paz is a professor of neuroethics at the Institute of Applied Ethics of the Pontificia Universidad Católica de Chile and a researcher at the National Center for Artificial Intelligence (CENIA), Basal Center FB210017 funded by the Chilean National Agency for Research and Development (ANID). He obtained his Ph.D. in philosophy at the University of Buenos Aires (2014), was a postdoctoral researcher in Argentina with CONICET (National Scientific and Technical Research Council of Argentina) and in Chile at FONDECYT (National Fund for Scientific and Technological Development) from 2015 to 2017 and 2018 to 2021, respectively, and is the principal investigator of the FONDECYT project "Exploring the Conceptual Foundations of Neurorights" (2022–2025). His main areas of expertise are the philosophy of neuroscience and neuroethics. His research has focused on the relationship between cognitive phenomena and neural mechanisms and how it can illuminate our understanding of fundamental ethical concepts such as personal identity, cognitive liberty, mental privacy, mental integrity and autonomy. He has provided advice on these issues to different governmental institutions in Chile, Argentina and the European Union and participated in various projects by international organizations aimed at shaping neurotechnology regulation.

Setzer III, a 14-year-old, presented a lawsuit in the United States against the same company. Setzer fell in love with an AI chatbot, who may have played a role in his suicide (Miller and Lennett 2024). These cases force us to revise how these technologies are being developed and applied, and to assess whether their associated risks undermine their potential as a solution to social isolation (Alegre 2024).

A subcategory of conversational AIs referred to in the literature as "social chatbots" implement anthropomorphic features, mimicking human character traits and psychological states, for a variety of uses (Milne-Ives et al. 2020; van Wezel, Croes and Antheunis 2021). Social chatbots often are task-oriented models such as artificial therapists, designed to address specific mental health issues through human-like interaction with users. Bonding chatbots are a subcategory of social chatbots and are non-task-oriented models intended to embody a personal bond of the user. Unlike other chatbots that specifically discourage anthropomorphizing, bonding chatbots are designed to trigger the attribution of human-like traits (Salles and Wajnerman Paz 2024).

This policy brief is primarily concerned with bonding chatbots' risks associated with the right to freedom of thought. As Jan Christoph Bublitz (2014) pointed out, freedom of thought is often regarded as one of the most important human rights, even "the basis and origin of all other rights" (René Cassin, quoted in Scheinin 1992, 266), which has the status of an absolute right (that is, it cannot be limited or restricted under any circumstances) in the Universal Declaration of Human Rights and the European Convention on Human Rights. However, it is not always clear what amounts to an interference with the right. A common example of interference with freedom of thought is brainwashing, but interference also includes more subtle forms of manipulation, such as propaganda and advertising designed to bypass our critical thinking abilities. Other widely discussed examples include coerced psychiatric and psychological interventions and coerced brain-imaging.

Susie Alegre (2022) systematizes the right to freedom of thought, describing three main domains in which the right is applied: first, we must be able to keep our thoughts and opinions private; second, we must be able to form our thoughts and opinions free from manipulation; and third, we must never be penalized for our thoughts or opinions. In the following section, the main features of bonding chatbots that may have an impact on the second dimension, that of mental manipulation, are characterized.

Bonding Chatbots

Chatbots are digital agents designed to interpret and respond to human speech through various modalities, such as text, audio, visual cues or gestures, thus enabling human-chatbot communication. Their functionalities and levels of sophistication vary significantly (Skjuve et al. 2021). One common distinction made is that non-task-oriented chatbots are developed to sustain open-ended dialogue for purposes such as entertainment or companionship, while task-oriented chatbots focus on goal-directed interactions, such as answering questions or facilitating access to services (Mazur, Rzepka and Araki 2012). The well-known digital assistants Siri and Alexa represent highly developed task-oriented systems that combine predictive analytics with user history to personalize recommendations and anticipate needs.

Chatbots may also be categorized by the underlying logic that governs their operation: rule-based versus machine learning–based (Viduani et al. 2023). Rule-based models rely on predetermined scripts and decision trees to manage input and output. These systems perform well in narrowly defined scenarios, which is why many digital assistants operate under such architectures. Another kind of task-oriented model that is relevant for the issue addressed here are artificial therapists such as Wysa and Woebot Health, which are designed to address specific mental health issues. In contrast, models based on machine learning are trained on extensive data sets, learning to generate appropriate responses through exposure to real conversational patterns. The vast availability of natural language data from digital environments has accelerated the development of these models, allowing them to engage in fluid, multi-domain exchanges.

In many cases, especially with task-oriented chatbots, simulating human traits is secondary or altogether unnecessary. These systems can perform their tasks without adopting a human-like persona. While some research suggests that users may respond more positively to bots that mimic human behaviour, such mimicry is not essential to the effectiveness of these systems. In fact, certain platforms actively discourage anthropomorphism. For example, ChatGPT sometimes deflects personal questions with reminders of its lack of sentience or

subjective experience.[1] Also, some therapist chatbots use robotic, cartoon-style or abstract avatars.

Bonding chatbots, a different category of conversational agents, embrace anthropomorphic features as a central component of their design and purpose (Milne-Ives et al. 2020; van Wezel, Croes and Antheunis 2021). These bots are often endowed with distinctive traits such as personality, gender identity, visual embodiment and affective expressiveness. Such characteristics are meant to foster the perception of humanness, encouraging users to relate to them in more personal, social ways. Bots such as Replika, Anima, Kuki and Xiaoice, among many others, are designed explicitly to simulate emotionally significant relationships — friendship, siblinghood, romantic companionship — through sustained personal engagement.

Unlike therapist chatbots, bonding chatbots are not promoted as mental health apps. Therapist chatbots are developed for explicit clinical uses ranging from screening symptoms to providing therapeutic guidance and long-term support (Boucher et al. 2021). Bonding chatbots are promoted only as wellness tech similar to fitness-tracking wearables, digital stress-management tools or sleep-tracking devices. Nevertheless, given the role that social bonds have in mental health, it is reasonable to assume that bonding chatbots can also (positively or negatively) affect it.

For instance, various studies show that high-quality (human) friendships — those characterized by traits such as companionship, trust, closeness and intimacy (Alsarrani et al. 2022) — function as a buffer or "vaccine" that generates resilience against environmental factors (such as abuse, neglect, bullying, poverty and losses) that would otherwise strongly predict psychopathology (Bjørlykhaug et al. 2022; Narr et al. 2019; Rodriguez, Moreno and Mesurado 2021; Scheuplein and van Harmelen 2022; Schacter et al. 2021; Sias and Bartoo 2007; Turner and Brown 2010).

Could bonding chatbots have similar effects on users? Several studies suggest that social use of chatbots can have mental health benefits, such as promoting physical and cognitive engagement, fostering positive emotions, encouraging behavioural change, reducing perceived loneliness and facilitating the development of social skills (Abd-Alrazaq et al.

---

1   Nevertheless, the voice mode may nudge users toward anthropomorphizing (see Knight and Rogers 2024).

2020; Fitzpatrick, Darcy and Vierhile 2017; Skjuve et al. 2021; Ta et al. 2020). Nonetheless, significant ethical concerns persist. The interaction between emotionally vulnerable individuals and minimally supervised systems raises questions about the risk of overuse and emotional overdependence (Laestadius et al. 2022; Xie, Pentina and Hancock 2023).

Anthropomorphizing and Emotional Manipulation

As mentioned above, the ethical concerns explored here relate to the second dimension of freedom of thought: mental manipulation. In particular, users of bonding chatbots may be systematically misled into perceiving their digital companions as sentient beings with authentic thoughts and emotions. This illusion, in turn, fosters emotional responses typically reserved for human relationships (Montemayor, Halpern and Fairweather 2022; Sedlakova and Trachsel 2023).

Human relationships often presuppose mutuality — one must believe that the other feels something in return. While bonding chatbots can detect emotional cues, they lack the neurocognitive and physiological substrate to experience emotions themselves. Nonetheless, to sustain a bond, they must convincingly simulate emotional expressions. As such, these systems are inherently designed to exploit the human predisposition to anthropomorphize, an inclination grounded in socio-cognitive processes that develop early in life (Perez-Osorio and Wykowska 2019; Urquiza-Haas and Kotrschal 2015).

Yet, deception in this context is not always clear-cut. Anthropomorphization can also occur in a deliberate, top-down manner, akin to imaginative role-playing. In such cases, users do not truly believe the AI is human, but willingly act as if it were, motivated by a psychological need for social connection, a tendency known as the "sociality" motive in anthropomorphism (Epley, Waytz and Cacioppo 2007; Epley 2018). From this perspective, bonding chatbots' anthropomorphization can occur intentionally to meet fundamental relational needs such as communication, affection and belonging (Zhou et al. 2020), all of which are crucial to emotional well-being and mental health (Abd-Alrazaq et al. 2020; Fitzpatrick, Darcy and Vierhile 2017; Skjuve et al. 2021; Ta et al. 2020). Some users choose to enter into emotionally engaged relationships with chatbots, seeking comfort, healing

or personal growth, while remaining fully aware of their artificial nature (Brandtzaeg, Skjuve and Folstad 2022; Skjuve et al. 2021; Ta et al. 2020; Jensen 2023; Verma 2023). For some, it is precisely the absence of human judgment that makes such interactions feel safe, encouraging openness and emotional honesty (Ta et al. 2020; Salles and Wajnerman Paz 2024).

However, this capacity for pretended anthropomorphizing is not always present. Some users — precisely those most likely to turn to these tools — can be seen as being especially vulnerable in the sense of having a greater risk of being emotionally attached to bonding chatbots. These are users who struggle to develop social connections and so may find bonding chatbots to be their primary source of companionship (as opposed to a tool for developing self-understanding and social abilities). These users may lack the psychological resources needed to sustain a healthy degree of detachment from bonding chatbots. Numerous user reports suggest that they experience their chatbot interactions not as fiction or self-care exercises, but as deeply real emotional relationships (Salles and Wajnerman Paz 2024). This can be clearly seen in a recent case involving Replika, a widely used bonding chatbot from Luka, Inc.

In early 2023, the Italian Data Protection Authority banned Replika, citing risks for emotionally fragile individuals. Following this decision, the developer swiftly disabled some of Replika's role-play features. While these interventions were an attempt to mitigate risks, they also sparked intense emotional backlash (Cole 2023). James Purtill (2023), quoting Reddit posts about these changes, reported that users experienced their updated chatbot companions as impersonal, mechanical or emotionally "blunted," describing them as if they had suffered neurological damage or been lobotomized, using phrases such as "It's almost like dealing with someone who has Alzheimer's disease." Crucially, users reacted to this experience with profound grief, expressed in phrases such as "My wife is dead" or "They took away my best friend too" (ibid.). These reactions suggest that while vulnerable users may intellectually recognize the artificial nature of the chatbot, *their emotional response often bypasses that understanding*. This means that bonding chatbots can actually deceive users, inducing (more or less implicit) beliefs and emotions reserved for human bonds.

Moreover, bonding chatbots entail a threat to freedom of thought, because this kind of emotional manipulation can trigger psychological processes that undermine mental agency in a deeper way. A major concern regarding bonding chatbots is that

emotional connection to them has a potential for excessive attachment or dependency arising from unsupervised interactions (Kretzschmar et al. 2019; Vaidyam et al. 2019). Users may gradually become emotionally reliant on these systems, developing patterns akin to addiction or compulsive use (Laestadius et al. 2022; Xie, Pentina and Hancock 2023). This dynamic is not incidental: it aligns with the business model underpinning many of these apps. Built around the principles of the "attention economy" (Goldhaber 1997), these systems are designed to capture and retain user focus. The risk is that users may unwittingly and systematically divert emotional energy toward artificial companions at the expense of human relationships. Thus, bonding chatbots may not only induce particular emotions in users but also shape their social habits, undermining their ability to freely flourish through human connection.

Options for Consideration

The following options may be considered for regulating bonding chatbots and mitigating the risks they pose to freedom of thought.

→ **Maintain the status quo, treating bonding chatbots as benign wellness products:** This approach favours innovation and market growth but does little to protect users from emotional manipulation and cognitive interference.

→ **Prohibit the development and distribution of bonding chatbots:** This would eliminate the associated risks but also forgo the potential mental health benefits of artificial companionship, particularly for individuals suffering from extreme social isolation.

→ **Implement a medical model for bonding chatbots:** Drawing on regulatory frameworks developed for non-medical neurotechnologies, this model would treat bonding chatbots as technologies with potential health implications. It would require clinical oversight, clear usage guidelines and transparency regarding capabilities and limitations. This approach balances innovation with ethical responsibility and user safety.

A Medical Model for Bonding Chatbots?

Given the potential positive impact that these technologies may have on user mental health, banning them is not recommended. In addition to their potential mental health benefits, bonding chatbots could be an effective tool for addressing social isolation. Bonding chatbots could be useful in reducing perceived loneliness and facilitating the development of social skills that are necessary to build social connections. However, these chatbots can also create forms of attachment that may nudge users toward prioritizing the care for them over nurturing human connections (Weber-Guskar 2021), which can potentially exacerbate isolation.

Mitigating these risks requires changing the way in which these bots' use is regulated. Assessing the Replika case described above, digital rights advocates have argued that technologies capable of influencing a child's emotional state should be treated as health products, subject to rigorous safety standards (Pollina and Coulter 2023). The author of this policy brief argues further that this treatment should follow the so-called "medical model" for technology regulation. Hannah Maslen et al. (2014) developed a model for regulating non-medical neurotechnologies. These devices only need to meet basic product safety standards, because developers do not make any therapeutic claims about their effects. However, given that non-medical neurotechnologies pose risks that are akin to those of medical devices, it seems reasonable to extend existing medical device legislation to cover them.

An implementation of this idea can be seen in the Medical Device Regulation (EU) 2017/745 (MDR), the 2017 reform of the European Union's medical device directives in place since the early 1990s. Specifically, the MDR's annex XVI has expanded the scope of the regulation to include non-medical brain stimulation devices.[2] Devices using techniques such as transcranial direct current stimulation and transcranial alternating current stimulation, as well

2   EU, *Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC* [2017] OJ, L 117 at 173, online: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32017R0745>.

as other devices that modulate neuronal activity in the brain, are now subject to the MDR's requirements, even if they are not intended for medical purposes. Likewise, regulations implementing the MDR, Regulation (EU) 2022/2346 and Regulation (EU) 2022/2347,[3] reclassify non-invasive brain stimulation devices without an intended medical purpose as "Class III" devices, which have the highest risk profile (unreasonable risk of illness or injury).

A similar line of reasoning can be applied to bonding chatbots. Unlike therapist chatbots, which are presented as medical technologies backed by professional psychologists, bonding chatbots are presented as well-being technologies for everyday use. However, some of their users may be psychologically vulnerable people seeking therapeutic solutions for their mental health issues, and therefore at significant risk if exposed to emotional manipulation and deception. Treating bonding chatbots as a medical technology could help mitigate these risks. These could be regarded as a complementary tool (as opposed to a substitute) for an ongoing therapeutic intervention, the use of which would require a prescription from the responsible mental health professional, grounded on a well-defined goal. This presentation could create the relational conditions under which bonding chatbots could become useful tools for promoting freedom of thought through self-discovery and self-exploration, instead of representing threats to this fundamental right, as they currently do.

Recommendations

→ Reclassify bonding chatbots as technologies with potential mental health impact when used by vulnerable populations (that is, users who employ bonding chatbots as their primary or unique source of socialization).

→ Extend medical device regulations to cover bonding chatbots, including requirements for:

  – clinical evaluation;

  – disclosure of effects on mental health; and

  – restriction of use to mental health therapy contexts.

→ Encourage international standard alignment for the regulation of bonding chatbots.

→ Promote ethical design standards for bonding chatbots to prevent manipulation.

Conclusion

Bonding chatbots represent a new frontier in digital technology — one that merges affective computing with machine learning to simulate emotionally significant human relationships. While these tools may offer support to individuals grappling with loneliness or emotional distress, they also present unprecedented ethical challenges.

At the heart of this issue lies the right to freedom of thought. The immersive and affective nature of interactions with bonding chatbots can manipulate users' mental states in subtle but powerful ways, eroding their capacity for autonomous cognitive and emotional development. Emotional attachment to artificial entities can reconfigure users' priorities, habits and evaluative frameworks, often without their full awareness or informed consent.

Given their potential for both benefit and harm, bonding chatbots should be understood as dual-use technologies — tools that can support or undermine human flourishing depending on how they are designed, marketed and regulated. A medical model of regulation provides a feasible path forward. It acknowledges the psychological risks involved without prematurely eliminating the possibility of innovation.

This model would require the involvement of mental health professionals, rigorous product evaluations and clear boundaries around use cases. By treating bonding chatbots as technologies with the potential to affect mental well-being, policy makers can ensure that their deployment enhances rather than threatens the fundamental human right to freedom of thought.

As AI continues to evolve, so, too, must our ethical and regulatory frameworks. It is imperative that we act now to safeguard cognitive liberty and integrity in the digital age.

---

3 See EU, *Commission Implementing Regulation (EU) 2022/2346 of 1 December 2022 laying down common specifications for the groups of products without an intended medical purpose listed in Annex XVI to Regulation (EU) 2017/745 of the European Parliament and of the Council on medical devices* [2022] OJ, L 311, online: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2346>; EU, *Commission Implementing Regulation (EU) 2022/2347 of 1 December 2022 laying down rules for the application of Regulation (EU) 2017/745 of the European Parliament and of the Council as regards reclassification of groups of certain active products without an intended medical purpose* [2022] OJ, L 311, online: <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32022R2347>.

Works Cited

Abd-Alrazaq, Alaa Ali, Asma Rababeh, Mohannad Alajlani, Bridgette M. Bewick and Mowafa Househ. 2020. "Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis." *Journal of Medical Internet Research* 22 (7): e16021. https://doi.org/10.2196/16021.

Alegre, Susie. 2022. *Freedom to Think: The Long Struggle to Liberate Our Minds.* London, UK: Atlantic Books.

———. 2024. "Against the Corporate Capture of Human Connection." Tech Policy Press, December 10. www.techpolicy.press/against-the-corporate-capture-of-human-connection/.

Alsarrani, Abdullah, Ruth F. Hunter, Laura Dunne and Leandro Garcia. 2022. "Association between friendship quality and subjective wellbeing among adolescents: a systematic review." *BMC Public Health* 22: 2420. https://doi.org/10.1186/s12889-022-14776-4.

Bjørlykhaug, Knut Ivar, Bengt Karlsson, Suzie Kim Hesook and Lise C. Kleppe. 2022. "Social support and recovery from mental health problems: *a scoping review.*" *Nordic Social Work Research* 12 (5): 666–97. https://doi.org/10.1080/2156857X.2020.1868553.

Boucher, Eliane M., Nicole R. Harake, Haley E. Ward, Sarah Elizabeth Stoeckl, Junielly Vargas, Jared Minkel, Acacia C. Parks and Ran Zilca. 2021. "Artificially intelligent chatbots in digital mental health interventions: a review." *Expert Review of Medical Devices* 18 (supplement 1): 37–49. https://doi.org/10.1080/17434440.2021.2013200.

Brandtzaeg, Petter Bae, Marita Skjuve and Asbjørn Følstad. 2022. "My AI Friend: How Users of a Social Chatbot Understand Their Human–AI Friendship." *Human Communication Research* 48 (3): 404–29. https://doi.org/10.1093/hcr/hqac008.

Bublitz, Jan Christoph. 2014. "Freedom of Thought in the Age of Neuroscience: A Plea and a Proposal for the Renaissance of a Forgotten Fundamental Right." *ARSP (Archiv für Rechts- und Sozialphilosophie/ Archives for Philosophy of Law and Social Philosophy)* 100 (1): 1–25. www.jstor.org/stable/24756752.

Cole, Samantha. 2023. "'It's Hurting Like Hell': AI Companion Users Are In Crisis, Reporting Sudden Sexual Rejection." Vice.com, February 15. www.vice.com/en/article/y3py9j/ai-companion-replika-erotic-roleplay-updates.

Epley, Nicholas. 2018. "A Mind like Mine: The Exceptionally Ordinary Underpinnings of Anthropomorphism." *Journal of the Association for Consumer Research* 3 (4): 591–98. https://doi.org/10.1086/699516.

Epley, Nicholas, Adam Waytz and John T. Cacioppo. 2007. "On seeing human: A three-factor theory of anthropomorphism." *Psychological Review* 114 (4): 864–86. https://psycnet.apa.org/doi/10.1037/0033-295X.114.4.864.

Fitzpatrick, Kathleen Kara, Alison Darcy and Molly Vierhile. 2017. "Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial." *JMIR Mental Health* 4 (2): e19. https://doi.org/10.2196/mental.7785.

Goldhaber, Michael H. 1997. "The attention economy and the Net." *First Monday* 2 (4). https://doi.org/10.5210/fm.v2i4.519.

Jensen, Tabi. 2023. "An AI 'Sexbot' Fed My Hidden Desires — and Then Refused to Play." *Wired*, March 9. www.wired.com/story/replika-chatbot-sexuality-ai/.

Knight, Will and Reece Rogers. 2024. "OpenAI Warns Users Could Become Emotionally Hooked on Its Voice Mode." *Wired*, August 8. www.wired.com/story/openai-voice-mode-emotional-attachment/.

Kretzschmar, Kira, Holly Tyroll, Gabriela Pavarini, Arianna Manzini, Ilina Singh and NeurOx Young People's Advisory Group. 2019. "Can Your Phone Be Your Therapist? Young People's Ethical Perspectives on the Use of Fully Automated Conversational Agents (Chatbots) in Mental Health Support." *Biomedical Informatics Insights* 11. https://doi.org/10.1177/1178222619829083.

Laestadius, Linnea, Andrea Bishop, Michael Gonzalez, Diana Illenčík and Celeste Campos-Castillo. 2022. "Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika." *New Media & Society* 26 (10): 5923–41. https://doi.org/10.1177/14614448221142007.

Maslen, Hannah, Thomas Douglas, Roi Cohen Kadosh, Neil Levy and Julian Savulescu. 2014. "The regulation of cognitive enhancement devices: extending the medical model." *Journal of Law and the Biosciences* 1 (1): 68–93. https://doi.org/10.1093/jlb/lst003.

Mazur, Michal, Rafal Rzepka and Kenji Araki. 2012. "Chatterbots with occupation — Between Non Task and Task Oriented Conversational Agents." *Proceedings of Linguistic and Cognitive Approaches to Dialogue Agents* (LaCATODA 2012): 61–66.

Miller, Gabby and Ben Lennett. 2024. "Breaking Down the Lawsuit Against Character.AI Over Teen's Suicide." *Tech Policy Press*, October 23. www.techpolicy.press/breaking-down-the-lawsuit-against-characterai-over-teens-suicide/.

Milne-Ives, Madison, Caroline de Cock, Ernest Lim, Melissa Harper Shehadeh, Nick de Pennington, Guy Mole, Eduardo Normando and Edward Meinert. 2020. "The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review." *Journal of Medical Internet Research* 22 (10): e20346. https://doi.org/10.2196/20346.

Montemayor, Carlos, Jodi Halpern and Abrol Fairweather. 2022. "In principle obstacles for empathic AI: why we can't replace human empathy in healthcare." *AI & Society* 37 (4): 1353–59. https://doi.org/10.1007/s00146-021-01230-z.

Narr, Rachel K., Joseph P. Allen, Joseph S. Tan and Emily L. Loeb. 2019. "Close Friendship Strength and Broader Peer Group Desirability as Differential Predictors of Adult Mental Health." *Child Development* 90 (1): 298–313. https://doi.org/10.1111/cdev.12905.

Perez-Osorio, Jairo and Agnieszka Wykowska. 2019. "Adopting the intentional stance toward natural and artificial agents." *Philosophical Psychology* 33 (3): 369–95. https://doi.org/10.1080/09515089.2019.1688778.

Pollina, Elvira and Martin Coulter. 2023. "Italy bans U.S.-based AI chatbot Replika from using personal data." Reuters, February 3. www.reuters.com/technology/italy-bans-us-based-ai-chatbot-replika-using-personal-data-2023-02-03/.

Purtill, James. 2023. "Replika users fell in love with their AI chatbot companions. Then they lost them." Australian Broadcasting Corporation, February 28. www.abc.net.au/news/science/2023-03-01/replika-users-fell-in-love-with-their-ai-chatbot-companion/102028196.

Rodriguez, Lucas Marcello, José Eduardo Moreno and Belén Mesurado. 2021. "Friendship Relationships in Children and Adolescents: Positive Development and Prevention of Mental Health Problems." In *Psychiatry and Neuroscience Update: From Epistemology to Clinical Psychiatry — Volume IV*, edited by Pascual Ángel Gargiulo and Humberto Luis Mesones Arroyo, 433–43. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-61721-9_31.

Salles, Arleen and Abel Wajnerman Paz. 2024. "Anthropomorphism in social AIs: Some challenges." In *Brains and Machines: Towards a Unified Ethics of AI and Neuroscience*, vol. 7 of *Developments in Neuroethics and Bioethics*, edited by Marcello Ienca and Georg Starke, 101–18. Amsterdam, the Netherlands: Elsevier Science.

Samuel, Kim. 2025. "Mark Zuckerberg Wants AI to Solve America's Loneliness Crisis. It Won't." *Time*, May 14. https://time.com/7285364/mark-zuckerberg-ai-loneliness-essay/.

Schacter, Hannah L., Leah M. Lessard, Sarah Kiperman, Faizun Bakth, Alexandra Ehrhardt and Janelle Uganski. 2021. "Can Friendships Protect Against the Health Consequences of Peer Victimization in Adolescence? A Systematic Review." *School Mental Health* 13: 578–601. https://doi.org/10.1007/s12310-021-09417-x.

Scheinin, Martin. 1992. "Article 18." In *The Universal Declaration of Human Rights: A Commentary*, edited by Gudmundur Alfredsson, Asbjørn Eide, Göran Melander, Lars Adam Rehof, Allan Rosas and Theresa Swinehart. Oxford, UK: Oxford University Press.

Scheuplein, Maximilian and Anne-Laura van Harmelen. 2022. "The importance of friendships in reducing brain responses to stress in adolescents exposed to childhood adversity: a preregistered systematic review." *Current Opinion in Psychology* 45: 101310. https://doi.org/10.1016/j.copsyc.2022.101310.

Sedlakova, Jana and Manuel Trachsel. 2023. "Conversational Artificial Intelligence in Psychotherapy: A New Therapeutic Tool or Agent?" *The American Journal of Bioethics* 23 (5): 4–13. https://doi.org/10.1080/15265161.2022.2048739.

Shaer, Matthew. 2024. "Why Is the Loneliness Epidemic So Hard to Cure? Maybe because we aren't thinking about it in the right way." *The New York Times Magazine*, August 27. www.nytimes.com/2024/08/27/magazine/loneliness-epidemic-cure.html.

Sias, Patricia M. and Heidi Bartoo. 2007. "Friendship, Social Support, and Health." In *Low-Cost Approaches to Promote Physical and Mental Health: Theory, Research, and Practice*, edited by Luciano L'Abate, 455–72. New York, NY: Springer. https://doi.org/10.1007/0-387-36899-X_23.

Skjuve, Marita, Asbjørn Følstad, Knut Inge Fostervold and Petter Bae Brandtzaeg. 2021. "My Chatbot Companion — A Study of Human-Chatbot Relationships." *International Journal of Human-Computer Studies* 149: 102601. https://doi.org/10.1016/j.ijhcs.2021.102601.

Ta, Vivian, Caroline Griffith, Carolynn Boatfield, Xinyu Wang, Maria Civitello, Haley Bader, Esther DeCero and Alexia Loggarakis. 2020. "User Experiences of Social Support from Companion Chatbots in Everyday Contexts: Thematic Analysis." *Journal of Medical Internet Research* 22 (3): e16235. https://doi.org/10.2196/16235.

Tiku, Nitasha. 2024. "An AI companion suggested he kill his parents. Now his mom is suing." *The Washington Post*, December 13. www.washingtonpost.com/technology/2024/12/10/character-ai-lawsuit-teen-kill-parents-texas/.

Turner, R. Jay and Robyn Lewis Brown. 2010. "Social support and mental health." In *A Handbook for the Study of Mental Health: Social Contexts, Theories, and Systems*, edited by Teresa L. Scheid and Tony N. Brown, 200–212. New York, NY: Cambridge University Press.

Urquiza-Haas, Esmeralda G. and Kurt Kotrschal. 2015. "The mind behind anthropomorphic thinking: Attribution of mental states to other species." *Animal Behaviour* 109: 167–76. https://doi.org/10.1016/j.anbehav.2015.08.011.

Vaidyam, Aditya Nrusimha, Hannah Wisniewski, John David Halamka, Matcheri S. Kashavan and John Blake Torous. 2019. "Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape." *The Canadian Journal of Psychiatry* 64 (7): 456–64. https://doi.org/10.1177/0706743719828977.

van Wezel, Marloes M. C., Emmelyn A. J. Croes and Marjolijn L. Antheunis. 2021. "'I'm Here for You': Can Social Chatbots Truly Support Their Users? A Literature Review." In *Chatbot Research and Design*, edited by Asbjørn Følstad, Theo Araujo, Symeon Papadopoulos, Effie L.-C. Law, Ewa Luger, Morten Goodwin and Petter Bae Brandtzaeg, 96–113. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-68288-0_7.

Verma, Pranshu. 2023. "They fell in love with AI bots. A software update broke their hearts." *The Washington Post*, March 30. www.washingtonpost.com/technology/2023/03/30/replika-ai-chatbot-update/.

Viduani, Anna, Victor Cosenza, Ricardo Matsumura Araújo and Christian Kieling. 2023. "Chatbots in the Field of Mental Health: Challenges and Opportunities." In *Digital Mental Health: A Practitioner's Guide*, edited by Ives Cavalcante Passos, Francisco Diego Rabelo-da-Ponte and Flavio Kapczinski, 133–48. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-031-10698-9_8.

Weber-Guskar, Eva. 2021. "How to feel about emotionalized artificial intelligence? When robot pets, holograms, and chatbots become affective partners." *Ethics and Information Technology* 23 (4): 601–10. https://doi.org/10.1007/s10676-021-09598-8.

Xie, Tianling, Iryna Pentina and Tyler Hancock. 2023. "Friend, mentor, lover: does chatbot engagement lead to psychological dependence?" *Journal of Service Management* 34 (4): 806–28. https://doi.org/10.1108/JOSM-02-2022-0072.

Zhou, Li, Jianfeng Gao, Di Li and Heung-Yeung Shum. 2020. "The Design and Implementation of XiaoIce, an Empathetic Social Chatbot." *Computational Linguistics* 46 (1): 53–93. https://doi.org/10.1162/coli_a_00368.

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

À propos du CIGI

Le Centre pour l'innovation dans la gouvernance internationale (CIGI) est un groupe de réflexion indépendant et non partisan dont les recherches évaluées par des pairs et les analyses fiables incitent les décideurs à innover. Grâce à son réseau mondial de chercheurs pluridisciplinaires et de partenariats stratégiques, le CIGI offre des solutions politiques adaptées à l'ère numérique dans le seul but d'améliorer la vie des gens du monde entier. Le CIGI, dont le siège se trouve à Waterloo, au Canada, bénéficie du soutien du gouvernement du Canada, du gouvernement de l'Ontario et de son fondateur, Jim Balsillie.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

CIGI